CLAIMS

What is claimed is:

1. A method of categorizing a plurality of new electronic documents into a set of categories, each of the categories containing a plurality of training set documents, by using a matrix representing document similarity that is derived by combining two or more measures of document similarity.

2. A method as recited in Claim 1, wherein the measures of document similarity include hyperlink similarity.

3. A method as recited in Claim 2, in which two documents among the plurality of documents are considered similar to each other when there is a link from one to the other, or when the two documents link to, or are linked to by, a set of other associated documents.

4. A method as recited in Claim 3, in which certain hyperlinks have greater or lesser similarity weight than other hyperlinks, based on other features of the links or their source or destination documents.

5. A method as recited in Claim 1, wherein the measures of document similarity include a similarity of text of the documents.

6. A method as recited in Claim 5, wherein two documents are considered similar based on a comparison of word vectors derived from the text of each of the two documents.

7. A method as recited in Claim 5, wherein text similarity is determined in part based upon weight values assigned to words of the text, and wherein certain words have greater or lesser weight than other words.

8. A method as recited in Claim 1, wherein the measures of document similarity include user click-through similarity.

9. A method as recited in Claim 8, wherein two documents are considered similar based on user click-through similarity when the documents are associated with similar patterns of user click behavior, selected from among frequency of clicks, click context, duration of viewing, proximity in time to other clicks, or proximity in context to other clicks.

10. A method as recited in Claim 1, wherein the measures of document similarity are derived from patterns detected in user viewing of the documents.

11. A method as recited in Claim 10, wherein the user viewing information is monitored by a web caching system and stored in a log.

12. A method as recited in Claim 10, wherein two documents are considered similar based on patterns of user viewing behavior, including frequency of viewing, viewing context, duration of viewing, proximity in time to other documents viewed by the same user, or similarity of patterns of viewing by all users.

13. A method as recited in Claim 1, wherein the measures of document similarity include URL similarity.

14. A method as recited in Claim 13, wherein two documents are considered similar if a URL of each document contains similar URL sub-components.

15. A method as recited in Claim 1, wherein the measures of document similarity include multimedia similarity.

16. A method as recited in Claim 15, wherein two documents are considered similar based on features derived from multimedia components linked to or contained by the documents.

17. A method as recited in Claim 1, wherein the combination of two or more measures of document similarity is achieved by taking the union of each of a plurality of graphs, each graph describing one of the measures of document similarity, to compute a combined graph that describes the combined document similarity.

18. A method as recited in Claim 1, wherein the combination of two or more measures of document similarity is achieved by taking the intersection of each of a plurality of graphs, each graph describing one of the measures of document similarity, to compute a combined graph that describes the combined document similarity.

19. A method as recited in Claim 1, further comprising the step of extracting structural information from the similarity matrix to obtain new documents supported by the set of training documents for each category.

20. A method as recited in Claim 19, wherein the structural information is obtained by optimizing an objective function.

21. A method as recited in Claim 19, wherein the structural information is obtained by only approximately optimizing an objective function.

22. A method as recited in Claim 21, wherein approximately optimizing the objective function comprises repeated application of a growth transformation.

23. A method as recited in Claim 19, further comprising the step of creating and storing a second matrix that represents an interim score for each document in each category.

24. A method as recited in Claim 19, further comprising the steps of, periodically as the matrix is being computed, normalizing rows of the matrix by normalizing within each document, across all categories, whereby the score for one document in a particular category will depend on the scores for that document in all other categories.

25. A method as recited in Claim 19, further comprising the steps of, periodically as the matrix is being computed, normalizing columns of the matrix by normalizing within

each category, across all documents, whereby the score for one document in a particular category depends on the scores for all other documents in that category.

26. A method as recited in Claim 1, in which the categories come from a manually defined taxonomy.

27. A method as recited in Claim 1, wherein the categories are derived from logs of user queries.

28. A method as recited in Claim 1, further comprising the steps of creating and storing the matrix using columns representing documents and rows representing user sessions, and wherein values of elements of the second matrix represent interest in a document shown by a particular user in a particular session.

29. A method as recited in Claim 1, further comprising the steps of creating and storing the matrix using columns representing user sessions and rows representing documents, and wherein values of elements of the second matrix represent interest in a document shown by a particular user in a particular session.

30. A method as recited in Claim 28, wherein the element values are computed as a function of a time that a user has spent viewing a document associated with each element.

30

31. A method as recited in Claim 28, further comprising the steps of creating and storing a second matrix representing a Similarity between pairs of documents i and j, wherein the second matrix is derived by comparing pairs of column vectors or row vectors, respectively i and j of the first matrix.

31

32. A method as recited in Claim 28, further comprising the steps of creating and storing a second matrix representing a Similarity between pairs of documents i and j, by finding pairs of documents i and j which have high interest values for a particular user in a particular session or period of time.

33. The method recited in Claim 1, further comprising the steps of:

identifying a category of a classification taxonomy of the hypertext system in which a first electronic document is presently classified; and

if a second electronic document is found to be highly Similar, storing information that classifies the second electronic document into the category.

34. A computer-readable medium carrying one or more sequences of instructions, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of categorizing a plurality of new electronic documents into a set of categories, each of the categories containing a plurality of training set documents, by using a matrix representing document similarity that is derived by combining two or more measures of document similarity.

35. A method of categorizing a plurality of new electronic documents for use in a hypertext search system, the method comprising the steps of:

creating and storing a set of categories for the documents;

creating and storing a matrix, in which rows and columns identify documents, and in which each element of the matrix stores a value that represents a similarity among a pair of documents associated with a row and column that intersect at the element;

deriving each matrix value by combining two or more measures of similarity that are obtained by analysis of the documents.

36. A method as recited in Claim 35, further comprising the steps of:

for each measure of document similarity, creating and storing a graph of links;

creating and storing a combined graph that combines the graphs and that represents a generalized similarity of the documents;

computing a generalized similarity value for a pair of documents based on the combined graph.

37. A method as recited in Claim 36, further comprising the steps of classifying unclassified documents into category nodes of a taxonomy structure associated with the hypertext search system based on the generalized similarity value in combination with a comparison of a set of pre-classified training set of documents with a set of unclassified documents, to carry out classification.